Classification of Data Center to Maximize Energy Utilization and Save Total Cost of Ownership

Article *in* International Review on Computers and Software · September 2012

CITATION
1

READS
67

4 authors, including:

Mueen Uddin
Effat University
61 PUBLICATIONS **436** CITATIONS
SEE PROFILE

Azizah Abdul Rahman
Universiti Teknologi Malaysia
146 PUBLICATIONS **465** CITATIONS
SEE PROFILE

Suhail Kazi
Universiti Teknologi Malaysia
21 PUBLICATIONS **67** CITATIONS
SEE PROFILE

Praise Worthy Prize

International Review on Computers and Software (I.RE.CO.S.), Vol. 7, N. 5
ISSN 1828-6003 September 2012

Classification of Data Center to Maximize Energy Utilization and Save Total Cost of Ownership

Mueen Uddin[1], Azizah Abdul Rahman[2], Suhail Kazi[3], Raed Alsaqour[4]

Abstract – *Multi-tier data centers have become a norm for hosting modern Internet applications as they provide a flexible, modular, scalable and high performance environment. However, these benefits come at a price of the economic dent incurred in powering and cooling these large hosting centers. Energyefficiency, performance, power utilization and environmental considerations become critical considerations in designingand implementing large, cluster-based multi-tier data centers for supporting a multitude of services. Tierperformance standards are key features for comparing the capabilities of a particular design topology against others or to compare group of data centers. This paper proposes to categorize data centers in four tier level categories depending on different parameters like performance, facilities, throughput, and subsystems in data center.*

The proposed technique helps to measure the performance and efficiency accurately and resourcefully. The classification helps technical and non-technical data center managers in identifying the anticipated performance of site infrastructure and design topologies equipped with latest data center standards to measure the performance and efficiency in terms of energy utilization and emission of greenhouse gases very hazardous for environmental sustainability and global warming.

Keywords: *Energy Efficient Data Center, Green IT, Tier Level Requirements, Virtualization, Tier Level Correlation*

I. Introduction

There has been an unprecedented increase on the level of concern regarding climate change and environmental sustainability. Businesses are under increasing pressure from customers, shareholders and users to propose legislative changes to improve their environmental credentials. Likewise, the environmental impact of Information Technology under the banner of "Green IT" has started being discussed by academia, media and government.

IT professionals are expected to play significant roles in bringing Green IT to organizations, provided they are prepared, have developed or developing necessary capabilities to lead and support sustainability initiatives [1].

The first step in greening the data centers is to baseline all the requirements to get the maximum value out of data center greening program. Now more than ever, energy efficiency seems to be on everyone's minds. Faced with concerns such as global warming and skyrocketing energy costs, more and more companies are considering if and how to increase efficiency [2], [3].

Green growth is about addressing climate change in an aggressive manner while, at the same time, making the green technologies and industries needed to combat it the driver of national economic growth. But it is also much more than that.

It entails a new social and civilization paradigm shift away from the business assumptions and lifestyles of the industrial age to a new path that satisfies

the need for economic growth, social and corporate responsibility, and the integrity of the environment [4]. Traditionally, environmentalism has been perceived to be at odds with economic prosperity. Environmental stewardship encompasses the notion of balancing current resource consumption with the resource requirements of future generations [5].

Gartner emphasizes that Information and communication technology (ICT) industry was responsible for about 2% of global CO_2 emissions almost equivalent to the aviation industry [6]. An EPA report presented to U.S congress in 2007 emphasizes that; current energy consumption in data centers is leading to an annual increase in the emission of CO_2 (greenhouse gases) from 42.8 million metric tons ($MMTCO_2$) in 2007 to 67.9 $MMTCO_2$ in 2011 [7]. Intense media coverage has raised the awareness of people around the globe for climate change and greenhouse gas effect on global warming. More and more customers are now considering the "green" aspect IT in selecting products and services. Besides the environmental concern, businesses have begun to face risks caused by being non-environmentally friendly.

Reduction of CO_2 footprints is an important problem that has to be addressed in order to facilitate further

Manuscript received and revised August 2012, accepted September 2012

advancements in computing systems. The survey results from InfoTech's (2008) international survey of Asia, Europe, USA and the rest of the world shows that more than 50% of the survey respondents were strongly concerned about global warming and its effect on climate change [8]. 144 nations signed andbegan implementing Kyoto accords, to reduce (GHS) emissions by 29%.

The complexities involved in planning, designing, and deploying today's critical production data center environments have increased exponentially in recent years. This is largely attributable to a growing demand for highly available operational frameworks capable of supporting high-density technology systems, *always-on* applications, and aggressive business-service delivery models.

Data centers are the nerve cells and building blocks of any IT business organization, providing services and capabilities of centralized storage, backups, recovery, management, networking and dissemination of data in which the mechanical, lighting, electrical and computing systems are designed to provide maximum services and processes [9]. Design of high performance, power-efficient, and dependable cluster based data centers has become an imperative requirement for meeting the demands of ever increasing businesses specially e-businesses in almost every sector of the economy ranging from academic institutions, government agencies, telecom industry and a myriad of business enterprises.

Large companies such as Google, Amazon, Akamai, AOL, and Microsoft use thousands of servers in their data center environment to handle high volume of traffic for providing customized (24x7) service to end users. Microsoft is adding 20000 servers monthly to their server farms to meet ever-growing demands of users and businesses [10]. A recent report shows that financial firms spend around 1.8 billion dollars annually on data centers for their businesses [11]. However, it has been observed that data centers contribute to a considerable portion of the overall delay for web-based services and this delay is likely to increase with more dynamic web contents. Poor response time has significant financial implications for many commercial applications. The power consumption of data centers is also drawing much attention, leading to the concept of Green Data Centers.

The green data center has moved from the theoretical to the realistic, with IT leaders being challenged to construct new data centers or retrofit existing ones, with energy saving features, sustainable materials and other environmental efficiencies. The green data center is an energy-efficient, dense computing ecosystem where:

1. Software technologies control data growth and shrink capacity demands
2. Managers use Service Level Agreements to manage energy usage
3. Energy efficient computing infrastructure optimizes performance and utilization levels
4. Physical plant is engineered for maximum energy efficiency

The ability to influence key legislative decisions, such as auctioning versus grandfathering allowances will enable companies to position for competitive advantage due to the significant asset value of the allowances [12].

A green data center is a repository for storage, management and dissemination of data in which mechanical, lighting, electrical and computing systems are designed for maximum energy efficiency and minimum environmental impact [13]. Data centers are one of the organizations where the Greening process should begin. They can be downsized to maximum capacity so as to make businesses Greener as done by Verizon wireless by cutting number of data centers from 10 to 3, saving $20 million. Green data center operations strategically aligns IT organization with, company objective to achieve greater corporate social responsibility.

Data centers in U.S. consumed 61 billion kilowatt-hour of electricity in 2006 at the cost of $4.5 billion. It is estimated that data centers power consumption will increase by 4% to 8% annually and is expected to reach 100 billion kWh by 2011 [14]. It is therefore becomes pertinent that future data center's design must focus on three critical parameters: high performance, power/thermal-efficiency and reliability. Like the energy costs, the energy use in data centers is also doubling every five years [15]. Considering delayed capital investments, there has been very little focus on the emphasis of energy efficiency as it becoming a key measure for operational effectiveness for large data centers or server farms [7].

In recent years, power has become one of the most important concerns for enterprise data centers hosting thousands of high-density servers and providing outsourced business critical IT services. A well-known approach to reducing power consumption is to transition the hardware components from high-power states to low-power states whenever performance allows. For example, a widely used power-efficient server design is to have run-time measurement and control of the desired application performance by adjusting the CPU power states using Dynamic Voltage and Frequency Scaling (DVFS).

While this approach can effectively reduce the dynamic power of the system, it cannot minimize the system leakage power for maximized power savings [16]. An important aspect for data center operators and managers is to meet the service level agreements (SLAs) required by customers and end users, such as response time and throughput. Service level agreements are the key performance metrics for customer service and are part of customer commitments. It is therefore, utmost to guarantee the SLAs of the applications while minimizing the power consumption of the data center.

Server virtualization strategies have become common these days and being implemented for proper resource sharing in data centers. Virtualization technologies such as VMware and Xen can consolidate applications previously running on multiple physical servers onto a smaller number of physical servers, effectively reducing

the power consumption of a data center by shutting down unused servers. Server Consolidation technique increases the utilization ratio of already installed servers to almost 50 % or even more with little overhead and cost [17]. More importantly, live migration [18] allows the movement of a virtual machine (VM) from one physical host to another with a reasonably short downtime[19].

This function makes it possible to use server consolidation as an online management approach, i.e., having run-time estimation of resource requirements of every VM and dynamically re-mapping VMs to physical servers using live migration. Internet service providers for hosting modern applications with dynamic Web contents are increasingly using multi-tier data centers. Typical three-tier architecture with front-end Web servers, middle-level application servers and back-end database servers provides a modular, flexible and scalable environment for Web hosting [20].

The multi-tier implementation has become the de facto industry standard for developing scalable client server applications. Such application tends to see dynamically varying workloads that contain long-term variations such as time of day effects as well as short-term fluctuations due to flash crowds [20]. The peak to mean ratio is typically high.

It results in low resource utilization if over provisioning is used to service peak workload. Dynamic variation of resources among those applications is required which not only provides sufficient resources to meet application performance goal but also prevents waste of resources caused by over provisioning. It creates a big challenge to provision resources efficiently to diverse components used by multitier applications with distinct resource requirement characteristics in a shared virtualized data center, while satisfying their performance goals under fluctuating workload and unpredictable component failures.

The data center industry has experienced several evolutions over the past 20 years. The performance of data center largely depends on the people involved in the design process. An assortment of variations in data center design are investigated and stated by many researchers. This has prompted the need for categorization of data center into tier levels to help specify the availability, reliability and performance efficiency of data center, so that energy efficient and green data centers should be implemented. Data center managers' needs to weigh both the cost and tier level in order to measure the performance efficiently and true cost and benefit analysis.

One of the major changes is the development of Tier based performance standards to help data center industry and business owners to categorize their data center according to design, performance and services they provide into different levels called tier levels. These standards provide quantifiable plateaus and are a basis for comparing the capabilities of a particular design topology against other designs as well as the associated site availability metrics for the various levels.

The Tier Level approach is the foundation used by a number of data center owners/users, consultants and design professionals in establishing a *design versus performance* ranking approach to today's data center projects.

The Tier classifications are created to consistently describe the site-level infrastructure required to sustain data center operations, not the characteristics of individual systems or subsystems. Data centers are dependent upon the successful and integrated operation of many separate site infrastructure subsystems. Every subsystem and system must be consistently deployed with the same site uptime objective to satisfy the distinctive Tier requirements. High levels of end-user availability may be attained through the integration of complex IT architectures and network configurations that take advantage of synchronous applications running on multiple sites.

The most critical decision making perspective data center owners and designers must consider, when making inevitable tradeoffs, is what effect the decision has on the life cycle integrated operation of the IT environment in data centers.

There is significant potential for energy efficiency improvements in data centers. Many technologies are either commercially available or will soon be available that could improve the energy efficiency of microprocessors, servers, storage devices, network equipment and infrastructure systems. Still, there are plenty of unexplored, reasonable opportunities to improve energy efficiency. Selection of efficient IT equipment and reducing mechanical infrastructure increases the energy efficiency. Improvements are possible and necessary at the level of the whole facility, i.e., the system level and at the level of individual components.

It is not possible to optimize data center components without considering the system as a whole, still it is true that efficient components are important for achieving an efficient facility; for instance, efficient servers generate less waste heat that reduces the burden on the cooling system [21].

II. Problem Background

Data centers industry is facing many challenges to provide up to date services to end users meeting service level agreements defined. The major issues being discussed everywhere is the availability of power resources to meet ever growing demands of businesses. On the other hand, this power consumption is also causing problem of emission of greenhouse gases, very hazardous for environmental health and global warming. Other issues like infrastructure and space challenges, cooling and cost issues, security challenges are some of the major hurdles in the development of environment friendly and cost effective green data centers.

Data center industry along with other organizations like Green Grid, Energy Star, EPA, VMware, IBM,

Microsoft, Oracle, Dell are trying to overcome some of the challenges and issues listed by providing many solutions like virtualization, server consolidation, cloud computing, energy management and most importantly development of metrics for measuring the performance of data center in terms of energy efficiency and CO_2 emissions.

To overcome some of the environmental issues, green IT spans many focus areas and activities, including power management; data center design, layout, and location; the use of biodegradable materials; regulatory compliance; green metrics and green labeling; carbon footprint assessment tools and methodologies; and environment-related risk mitigation. A growing number of IT vendors and users have begun to turn their attention toward green IT, triggered by the imminent introduction of more green taxes and regulations; there will be a major increase in demand for green IT products and solutions. Green IT will be the hot topic for years to come, because it now becomes imperative to develop environmentally sustainable IT, from both an economic and an environmental viewpoint [2].

The infrastructure challenges such as determining data center architectures and providing sufficient cooling and power for large numbers of servers requires the classification of data center into some standardize types or tiers to categorize the data center into measureable units so that performance can be measured individually or collectively and some benchmarking standards should be set to be followed by data center managers to achieve energy efficient and green data centers. It is very much apparent that if we cannot measure the efficiency and performance of data center then how we can estimate the power requirements needed by data centers and availability of other resource equipments needed to fulfill end user requirements.

A popular trend currently being implemented in data center architecture is the use of large scale, modular data centers composed of shipping containers filled with servers [22], but more radical proposals range from micro data centers placed inside condominium closets [23]to floating barges filled with servers running off of power generated from ocean currents [24]. The massive scale of data centers has led to new distributed application architectures. Clustering of web servers and databases becomes necessary when a single commodity server cannot meet customer demands [25]. Large-scale data mining is also an increasingly popular use for data centers, with search engines becoming some of the largest consumers of data center resources. These systems employ clustering frameworks like MapReduce and Dryad to distribute work across many hundreds or thousands of nodes [26].

As data centers attempt to improve resource utilization through server consolidation, it also becomes necessary for data center operators to understand how the placement of applications impacts performance and resource consumption [27].Efficient resource management is a key concern for data center operators looking to both meet application Service Level Agreements and reduce costs. Shared hosting platforms attempt to multiplex physical resources between multiple customer applications [27]. Reliability becomes an important concern when running mission critical applications within data centers. The large-scale modern data center means that hardware components fail on a constant basis, requiring both low level fault tolerance techniques like RAID, and high-level reliability mechanisms within applications.

The increasing energy consumption by data centers is a growing concern; research is being done to overcome energy issues, especially in developing countries like Pakistan, which is already facing huge energy deficits even to provide electricity for domestic purposes. Many industries are closing down and there is a continuous decline in the economy of the country, due to the shortage of electricity and other forms of energy.

III. Data Center Performance Evaluation Considerations

Data center performance is now becoming a wide industry requirement to be achieved by implementing different strategies like considering green initiatives; server virtualization and green metrics to reduce the consumption of energy utilization and increase the availability and utilization of already installed servers and other devices. Energy efficiency in data centers is achieved by optimizing computing resource usage, by using the smallest computing resources to process maximum number of valuable tasks; it results in consuming smallest amount of energy to process maximum number of tasks.

In a homogeneous system, there is no difference between optimizing energy and optimizing the computing resource usage. But in the heterogeneous system, a solution of optimizing the computing resource usage may not energy efficient [28]. The Tier classification describes the site level infrastructure topology required to sustain data center operations, not the characteristics of individual systems or subsystems.

This standard is predicated on the fact that data centers are dependent upon the successful and integrated operation of several separate site infrastructure subsystems, the number of which is dependent upon the individual technologies (e.g., power generation, refrigeration, uninterruptible power sources.) selected to sustain the operation. Every subsystem and system integrated into the data center site infrastructure must be consistently deployed with the same site uptime objective to satisfy the distinctive tier requirements.The requirements of each tier is measured by outcome based confirmation tests and operational impacts. This method of measurement differs from a prescriptive design approach or a checklist of required equipment.

The data center tier performance level is determined and based on an evaluation and rating of 16 critical site infrastructure subsystems as compared against the

desired tier classification for the data center. Data centers are dependent upon the successful operation of these 16 subsystems shown in Table I. These systems and subsystems must be installed with the same availability objective to meet the tier level requirements and performance expectations.

TABLE I
ASPECTS CONSIDERED IN TIER PERFORMANCE EVALUATION

Category	Subsystem
Electrical	Utility Service
	Lightning Protection
	Power Backbone
	UPS Systems
	UPS Batteries
	Engine Generator
	Load Bank
	Critical Power Distribution
	Grounding
Mechanical	Raised Floor Cooling
	UPS Cooling
	Mechanical Plant
Support Systems	Contamination
	Fire Detection and Protection
	Physical Security
	Alarms and Monitoring

IV. Correlation of Tier Levels

The tier levels (Fig. 1) are closely correlated with each other as they share most of the functionalities and the components installed are almost same except the availability of services provided. Tier 4 data centers provide the most extensive services with high availability rate up to almost 99.99%. The Electrical, mechanical and plumbing aspects of data center are well defined in these tier levels to highlight the importance of individual components installed and used in data centers to provide services to end users.

These components are well defined in order to measure the efficiency of individual components and then cumulative efficiency of whole data center to highlight the performance of data center.

Energy and power efficiency are nowadays hot issues in data center industry being discussed by all academia, researchers.

It is therefore pertinent to highlight the relationship among individual components and then identifying suitable metrics to measure the performance and benchmark values to be used by data center managers. The correlation of tier level helps to measure these performances in terms of energy efficiency and CO2 emissions.

The four tier standard classifications address topology, or configuration, of site infrastructure, rather than a prescriptive list of components to achieve a desired operational outcome.

For example, the same number of chillers and UPS modules can be arranged on single power and cooling distribution paths resulting in a Tier 2 solution (Redundant Components), or on two distribution paths that may result in a Tier 3solution (Concurrently Maintainable).

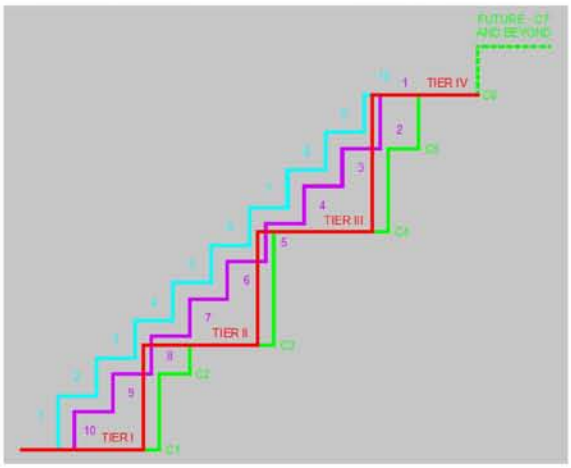

Fig. 1. Correlation of Various Tier Levels

V. Data Center Tier Requirements

The Tier classification requirements are listed in Table II, which indicate the preceding requirements for defining the four distinct Tier classification levels provided, as basis for comparing or describing the functionality, capacity, and cost of a data center's overall architecture.

These requirements should be considered as benchmarks for defining the performance of data center in terms of overall efficiency. These requirements must focus on the availability of the entire data center facility including power, connectivity and cooling components.

They should also describe Tier classifications as the degree to which the facility is resilient to failures of mechanical, electrical and plumbing (MEP) systems.

TABLE II
SUMMARY OF PRECEDING REQUIREMENTS DEFINING THE FOUR DISTINCT TIER CLASSIFICATION LEVELS

Requirements	Tier 1	Tier2	Tier3	Tier4
Adaptive capacity components to support IT Load	N	N+1	N+1	N after any failure
Distribution Paths	1	1	1 Active & 1 Alternate	2 Simultaneous active
Concurrently Maintainable	No	No	Yes	Yes
Fault Tolerance	No	No	No	Yes
Compartmentalization	No	No	No	Yes
Continuous Cooling	Load Density Dependent	Load Density Dependent	Load Density Dependent	Class A

VI. The Tier Classification Systems for Data Centers

The tier classification system measures the performance of data center operating infrastructure, which includes power, cooling, emergency backup, and fire suppression. The power and cooling capabilities of a facility are delivered by its MEP infrastructures.

The mechanical systems provide cooling to the

environment in which the data processing equipment is installed. These systems are composed of air handlers, air conditioners, chillers, plenums to channel airflow, and so on. The electrical systems provide the power to the data processing equipment. These systems are composed of the utility service to the facility, transfer switches, generators and Uninterruptible Power Supplies (UPS), batteries, Power Distribution Units (PDUs), load banks, breaker panels, copper cabling, etc. The plumbing systems support the mechanical and electrical systems by routing cabling, air, water, fire suppression gases, and so on. There are multiple plumbing circuits in a facility; these are analogous to the vascular system of the human body. It refers to the degree of resilience the data center has to failures of its MEP systems. Redundancy and topology of the infrastructure's design provide resilience to failures. In the tier classification model, a Tier facility is the *least resilient* and a Tier-4 is the "most resilient."

Therefore, a Tier-1 type data center has the lowest availability and a Tier-4 has the highest availability. This classification model also provides an academic and objective benchmark that helps describe and compare the functionality, capacity, and cost of data center infrastructures. At times, the drive to align the uptime of the IT facility with the business becomes bogged down in focusing on tier levels. Other factors beyond tier level compliance can impact uptime performance.

The following is a summary of representative site availability expectations for each of the tier levels described above. The availability percentages can be considered characteristic of the operating experiences of a representative number of sites within each tier classification:

- Tier 1= 28.8 hours and 99.67%
- Tier 2 = 22.0 hours and 99.75%
- Tier 3 = 1.6 hours and 99.98%
- Tier 4 = 0.4 hours and 99.99%

Telecommunication Industry Association started to work on the classification of data centers and in 2005.It started the project TIA-942 that categorizes the data centers into four categories [29]. As a rule, the overall Tier Level is based on the lowest tier ranking or weakest component. For example a data center may be rated tier 3 for electrical, but tier 2 for mechanical. The data centers overall tier rating is 2. In practice a data center may have different tier ratings for different portions of the infrastructure. These tier levels are:

1. Tier1: Basic Site Infrastructure
2. Tier2: Redundant Site Infrastructure Capacity Components
3. Tier3: Concurrently Maintainable Site Infrastructure
4. Tier4: Fault Tolerant Site Infrastructure

VI.1. Tier 1: Basic Site Infrastructured

Tier 1 data center facilities have no redundant capacity components. It provides basic power and cooling with no excess capacity for backup or failover. There is no redundancy in the MEP distribution paths. The unplanned outage or failure of a capacity component or distribution element will impact systems and customers. Maintenance needed for the MEP infrastructure to replace components or do utility work impacts the facility just as if there were an unplanned outage.

Tier 1 data centers typically experience two separate 12 hours site wide shutdowns per year for repair work. Additionally they also typically experience 1.2 equipment or distribution component failures on average each year. This equates to 28.8 hours of downtime per year, or 99.67% availability.

They acknowledge the owner's desire for dedicated site infrastructure to support IT systems. These infrastructures also provides an improved environment over that of an ordinary office setting and includes: a dedicated space for IT systems; a UPS to filter power spikes, sags, and momentary outages; dedicated cooling equipment not shut down at the end of normal office hours; and an engine generator to protect IT functions from extended power outages. Tier 1 data center may be suitable for small businesses where IT is intended for internal business processes.

VI.1.1. Features of Tier 1

The fundamental requirement

Tier 1 is a basic and simple data center having non-redundant capacity components and a single non-redundant distribution path serving the computer equipments like servers.

The performance confirmation tests

Tier 1 data centers have sufficient capacity to meet the needs of the site and provide all types of services to end-users. Planned work requires most or all of the site infrastructure systems to be shut down affecting computer equipment, systems, and end users. The availability is not achieved as required by end users and businesses.

The operational impacts

The site is susceptible to disruption from both planned and unplanned activities. Operation (Human) errors of site infrastructure components will cause a data center disruption.

An unplanned outage or failure of any capacity system, capacity component, or distribution element will impact the computer equipment.

The site infrastructure must be completely shut down on an annual basis to safely perform necessary preventive maintenance and repair work. Urgent situations may require more frequent shutdowns. Failure to regularly perform maintenance significantly increases the risk of unplanned disruption as well as the severity of the consequential failure.

VI.2. Tier 2: Redundant Data Center

Tier 2 data center has redundant capacity components,

but only a single non-redundant distribution path serving the data processing equipments. The benefit of this level is that any redundant capacity component can be removed from service on a planned basis without causing the data processing to be shut down. Tier 2 sites have average one unplanned outage per year, and scheduled three maintenance activities over a two-year period. The annual impact to operations is 22 hours of downtime per year, or 99.75% availability.

Tier II solutions include redundant critical power and cooling capacity components to provide an increased margin of safety against IT process disruptions due to site infrastructure equipment failures. The redundant components are typically extra UPS modules, chillers, heat rejection equipment, pumps, cooling units, and engine generators.

A malfunction or normal maintenance will result in loss of a capacity component. Tier 2 data center may be appropriate for internet-based companies without serious financial penalties for quality of service commitments

VI.2.1. Features of Tier 2

The fundamental requirement

Tier 2 data center has redundant capacity components and a single, non-redundant distribution path serving the computer equipment installed.

The performance confirmation tests

Redundant capacity components can be removed from processing they are performing on a planned basis without causing any of the computer equipment to be shut down.

Removing distribution paths for maintenance or other activity requires shutdown of computer equipment.

The operational impacts

Tier 2 data center is susceptible to disruption from both planned activities and unplanned events. Operation (Human) errors of site infrastructure components may cause a data center disruption.

An unplanned capacity component failure may impact the computer equipment. An unplanned outage or failure of any capacity system or distribution element will impact the computer equipment.

The site infrastructure must be completely shut down on an annual basis to safely perform preventive maintenance and repair work. Urgent situations may require more frequent shutdowns. Failure to regularly perform maintenance significantly increases the risk of unplanned disruption as well as the severity of the consequential failure.

VI.3. Tier 3: Redundant Data Center with Concurrent Maintenance

Tier 3 data center has redundant capacity components and multiple independent distribution paths serving the data processing footprint. There is sufficient MEP capacity to meet the needs of the data processing systems even when one of these redundant MEP components has been removed from the infrastructure.

Tier 3 data center can support maintenance activities and some unplanned events without interruption to the computing systems. Because of concurrent maintenance capability provided, no annual shutdowns for routine maintenance are required. These data centers have unplanned events totaling only 1.6 hours per year and delivers 99.98% availability. These infrastructures add the concept of Concurrent Maintenance beyond what is available in Tier 1 and Tier 2 solutions.

Concurrent Maintenance means that each and every capacity or distribution component necessary to support the IT processing environment can be maintained on a planned basis without impact to the IT environment. The effect on the site infrastructure topology is that a redundant delivery path for power and cooling is added to the redundant critical components of Tier 2.

Maintenance allows the equipment and distribution paths to be returned to like new condition on a frequent and regular basis. Thus, the system will reliably and predictably perform as originally intended. Moreover, the ability to concurrently allow site infrastructure maintenance and IT operation requires that each and every system or component that supports IT operations must be able to be taken offline for scheduled maintenance without impact to the IT environment. This concept extends to important subsystems such as control systems for the mechanical plant, start systems for engine generators, EPO controls, power sources for cooling equipment and pumps, isolation valves, and others. Tier 3application would include companies that span multiple time zones or whose information technology resources support automated business process.

VI.3.1. Features of Tier 3

The fundamental requirements

A Concurrently Maintainable data center has redundant capacity components and multiple independent distribution paths serving the computer equipments. Only one distribution path is required to serve the installed computer equipments at any time like servers and storage devices.

All IT equipment is dual powered as defined by the Institute's Fault Tolerant Power Compliance Specification, Version 2.0 and installed properly to be compatible with the topology of the site's architecture. Transfer devices, such as point-of-use switches, must be incorporated for computer equipment that does not meet this specification.

The performance confirmation tests

Each and every capacity component and element in the distribution paths can be removed from service on a planned basis without impacting any of the computer equipments. There is sufficient permanently installed capacity to meet the needs of the site when redundant

components are removed from service for any reason.

The operational impacts

The site is susceptible to disruption from unplanned activities. Operation errors of site infrastructure components may cause a computer disruption.

An unplanned outage or failure of any capacity system or component or distribution element will impact the computer equipment.

Using the redundant capacity components and distribution paths to safely work on the remaining equipments can perform planned site infrastructure maintenance.

During maintenance activities, the risk of disruption may be elevated. (This maintenance condition does not defeat the Tier rating achieved in normal operations).

VI.4. Tier 4: Fault Tolerant Site Infrastructure

Tier 4 data centers have multiple, independent, and physically separate systems, each having redundant capacity components and multiple, independent, diverse and active distribution paths supporting all data processing. There is no impact of any single point of failure of MEP component and distribution path has no negative impact to the data processing systems.

The impact of the data processing equipment failures is statistically reduced to 0.8 hour per year yielding 99.99% availability. Tier 4 infrastructures are built on Tier III, adding the concept of Fault Tolerance to the site infrastructure topology. Similar to the application of Concurrent Maintenance concepts, Fault Tolerance extends to each and every system or component that supports IT operations. The definition of Fault Tolerance in Tier 4 data centers is based on a single component or path failure. However, the site must be designed and operated to tolerate the cumulative impact of every site infrastructure component, system, and distribution path disrupted by the failure. For example, the failure of a single switchboard will affect every subpanel and equipment component deriving power from the switchboard.

A Tier 4 data center facility must tolerate these cumulative impacts without affecting the operation of the computer room. Tier 4 level data centers include companies who have extremely high-availability requirements for ongoing business such as E-commerce, market transactions, or financial settlement processes.

VI.4.1. Features of Tier 4

The fundamental requirements

A Fault Tolerant data center has multiple, independent, physically isolated systems that provide redundant capacity components and multiple, independent, diverse, active distribution paths simultaneously serving the computer equipment. The redundant capacity components and diverse distribution paths shall be configured such that "N" capacity is providing power and cooling to the computer equipment after any infrastructure failure.

All IT equipments are dual powered as defined by the Fault Tolerant Power Compliance Specification, Version 2.0 and installed properly to be compatible with the topology of the site's architecture. Transfer devices, such as point of use switches, must be incorporated for computer equipment that does not meet this specification.

Complementary systems and distribution paths must be physically isolated from one another (compartmentalized) to prevent any single event from simultaneously impacting both systems and distribution paths.

Continuous Cooling is required for achieving continuous Availability.

The performance confirmation tests

A single point of failure of any capacity system, capacity component, or distribution element will not impact the computer equipment. The system itself automatically responds *self-heals* to a failure to prevent further impact to the site.

Each and every capacity component and element in the distribution paths can be removed from service on a planned basis without impacting any of the computer equipments. There is sufficient capacity to meet the needs of the site when redundant components or distribution paths are removed from service for any reason.

The operational impacts

The data center is not susceptible to disruption from a single planned or unplanned event or activities.

Using the redundant capacity components and distribution paths to safely work on the remaining equipments can perform the site infrastructure maintenance.

During maintenance activity where redundant capacity components or a distribution path shut down, the computer equipment is exposed to an increased risk of disruption in the event a failure occurs on the remaining path. This maintenance configuration does not defeat the Tier rating achieved in normal operations.

Operation of the fire alarm, fire suppression, or the emergency power off (EPO) feature may cause a data center disruption.

VII. Conclusion

To accommodate increasingly dense technology environments, increasingly critical business applications, and increasingly stringent service level demands, data centers are typically engineered to deliver the highest-affordable availability levels facility-wide.

Within this monolithic design approach, the same levels of mechanical, electrical, and IT infrastructure are installed to support systems and applications regardless of their criticality or business risk if unplanned downtime occurs.

Typically, high redundancy designs are deployed in order to provide for all eventualities.

The result, in many instances, is to unnecessarily drive up either upfront construction or retrofitting costs and ongoing operating expenses. In this paper, we presented a novel multi-level tier approach for increasing the energy efficiency and maintaining the performance bounds of a multi-tier data centers.

The Tier Performance Standards are an owner/user set of requirements used to clearly define expectations for the design and management of the data center to meet a prescribed level of availability. The Tier Level Classification system is the foundation used by many data center owners/users, consultants and design professionals in establishing a *design versus performance* ranking approach to today's data center projects.

Acknowledgements

This research was supported in part by Ministry of Higher Education (MOHE), University Kebangsaan Malaysia (UKM), Malaysia, under Fundamental Research Grant Scheme (FRGS). Grant: FRGS/1/2012/SG05/UKM/02/7.

References

[1] A. Molla, V. A. Cooper, and S. Pittayachawan, "IT and eco-sustainability: Developing and validating a green IT readiness model," *ICIS 2009 Proceedings*, 2009.

[2] S. Murugesan, "Making IT green," *IT Professional, vol. 12*, 2010, pp. 4-5.

[3] J. Mikulik and M. Babina, "The Role of Universities in Environmental Management," *Polish Journal of Environmental Studies, vol. 18*, 2009, pp. 527-531.

[4] L. M. Bak, H. L. Solis, and G. Kleisterlee. (2010). *Green economy making.* Available: http://www.unep.org/pdf/OP_Feb/EN/OP-2010-02-EN-FULLVERSION.pdf

[5] B. K. Sovacool and M. A. Brown, "Competing dimensions of energy security: An international perspective," *Annual Review of Environment and Resources, vol. 35*, 2010, pp. 77-108.

[6] L. Norford, A. Hatcher, J. Harris, J. Roturier, and O. Yu, "Electricity use in information technologies," *Annual Review of Energy, vol. 15*, 1990, pp. 423-453.

[7] R. Brown, E. Masanet, B. Nordman, B. Tschudi, A. Shehabi, J. Stanley, J. Koomey, D. Sartor, P. Chan, and J. Loper, "Report to congress on server and data center energy efficiency," *Public law, vol. 109*, 2007, p. 431.

[8] I.-T. R. Group, "If You Measure It, They Will Green: Data Center Energy Efficiency Metrics," *Info-Tech Research Group, London*, 2008.

[9] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the art of virtualization," in *Proceedings of the nineteenth ACM symposium on Operating systems principles*, 2003, pp. 164-177.

[10] Microsoft, "How Microsoft Designs the Virtualization Host and Network Infrastructure," ed: Microsoft IT Showcase-Technical Case Study, 2009.

[11] FinanceTech. (2009). *Finance Technology Network.* Available: http://www.financetech.com

[12] U. Mueen, A. R. Azizah, and M. Jamshed, "Carbon sustainability framework to reduce CO2 emissions in data centres," *International Journal of Green Economics, vol. 5*, 2011, pp. 353-369.

[13] T. Daim, J. Justice, M. Krampits, M. Letts, G. Subramanian, and M. Thirumalai, "Data center metrics: An energy efficiency model for information technology managers," *Management of Environmental Quality: An International Journal, vol. 20*, 2009, pp. 712-731.

[14] M. Marwah, R. Sharma, R. Shih, C. Patel, V. Bhatia, M. Mekanapurath, R. Velumani, and S. Velayudhan, "Data analysis, visualization and knowledge discovery in sustainable data centers," in *COMPUTE '09 Proceedings of the 2nd Bangalore Annual Compute Conference*, 2009, p. 2.

[15] Gartner, "Sustainable IT," ed: A Gartner Briefing, Gartner, Dublin, 2008.

[16] Y. Wang, X. Wang, M. Chen, and X. Zhu, "Power-efficient response time guarantees for virtualized enterprise servers," in *Real-Time Systems Symposium*, 2008, pp. 303-312.

[17] M. Uddin and A. A. Rahman, "Server Consolidation: An Approach to make Data Centers Energy Efficient and Green," *International journal of Scientific and Engineering Research, (IJSER), vol. 1*, 2010, pp. 1-7.

[18] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live migration of virtual machines," in *NSDI'05 Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation*, 2005, pp. 273-286.

[19] M. R. Hines and K. Gopalan, "Post-copy based live virtual machine migration using adaptive pre-paging and dynamic self-ballooning," in *VEE '09 Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments*, 2009, pp. 51-60.

[20] B. Urgaonkar, P. Shenoy, A. Chandra, and P. Goyal, "Dynamic provisioning of multi-tier internet applications," in *Proceedings of 2nd International Conference on Autonomic Computing*, 2005, pp. 217-228.

[21] M. Uddin and A. A. Rahman, "Techniques to implement in green data centres to achieve energy efficiency and reduce global warming effects," *International Journal of Global Warming, vol. 3*, 2011, pp. 372-389.

[22] R. H. Katz, "Tech titans building boom," *Spectrum, IEEE, vol. 46*, 2009, pp. 40-54.

[23] K. Church, A. Greenberg, and J. Hamilton, "On delivering embarrassingly distributed cloud services," *Hotnets VII, vol. 34*, 2008.

[24] J. Clidaras and D. W. Stiver, "Water-based data center," ed: WO Patent WO/2010/129,341, 2010.

[25] R. Levy, J. Nagarajarao, G. Pacifici, A. Spreitzer, A. Tantawi, and A. Youssef, "Performance management for cluster based web services," in *IFIP/IEEE Eighth International Symposium on Integrated Network Management*, 2003, pp. 247-261.

[26] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: distributed data-parallel programs from sequential building blocks," in *EuroSys '07 Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007*, 2007, pp. 59-72.

[27] A. Chandra, W. Gong, and P. Shenoy, "Dynamic resource allocation for shared data centers using online measurements," in *Proceedings of the ACM SIGMETRICS international conference on Measurement and modelling of computer systems*, 2003, pp. 381-398.

[28] R. Van den Bossche, K. Vanmechelen, and J. Broeckhove, "An evaluation of the benefits of fine-grained value-based scheduling on general purpose clusters," *Future Generation Computer Systems, vol. 27*, 2011, pp. 1-9.

[29] T. I. Association, "ANSI/TIA 942—2005 《 Telecommunications Infrastructure Standard for Data Centers," ed: USA: TELECOMMUNICATIONS INDUSTRY ASSOCIATION Standards and Technology Department.

Authors' information

[1]Faculty of Computing and Technology,Asia Pacific University of Technology & Innovation, Bukit Jalil, 57000, Kuala Lumpur, Malaysia.

[1,2]Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310,Skudai, Johor, Malaysia.

[3]Faculty of Mechanical Engineering, Universiti Teknologi Malaysia, 81310,Skudai, Johor, Malaysia.

[4]School of Computer Science, Faculty of Information Science andTechnology, University Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.

Mueen Uddin is a Senior Lecturer at Faculty of Computing &Technology, Asia Pacific University of Technology & Innovation. He completed his PhD in Information Systems from UTM Malaysia in 2012, BS & MS in Computer Science from Isra University Hyderabad Pakistan in 2008 with specialty in Information networks. His research interests include Green IT, energy efficient data centers and Virtualization technologies, digital content protection and deep packet inspection, intrusion detection and prevention systems, MANET routing protocols and their analysis. Dr. Mueen has over 22 international Journal publications with many Conference papers.

Azizah Abdul Rahman is an Associate Professor at Faculty of Computer Science and Information Systems, University Teknologi Malaysia. She completed her B.Sc and M.Sc from USA, and PhD in information systems from University Technology Malaysia. Her research interests include designing and implementing techniques for information systems in an organizational perspective, knowledge management, designing networking systems in reconfigurable hardware and software, and implementing security protocols needed for E-businesses. Dr Azizah has more than 40 publications in international journals and Conferences in the field of Green IT, Knowledge management, information systems design and implementation.

Suhail Kazi is a Senior Lecturer at Faculty of Mechanical Engineering, University Teknologi Malaysia. He obtained his B.Sc in Mechanical Engineering from Mehran University of Engineering and Technology, Jamshoro, Pakistan in 1999. He then pursed his M.Sc in Computer Science from Isra University, Hyderabad, Pakistan in 2005. He completed his PhD at Universiti Teknologi Malaysia in 2011. His major field of Study is active tremor control and mechatronics. His research interests include intelligent active force control, Energy harvesting, Vibration harvesting, system modeling and simulation, vibration control and mechatronics.

Raed Alsaqour is an Assistant Professor in the School of Computer Science, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia. He received his B.Sc degree in computer science from Mu'tah University, Jordan, in 1997. M. Scdegree in distributed system from University Putra Malaysia, Malaysia, in 2003 and his PhD degree in wireless communication system from Universiti Kebangsaan Malaysia, Malaysia, in 2008. His research interests include wireless network, ad hoc network, vehicular network, routing protocols, simulation, network performance evaluation, and Green IT. He also has a keen interest in computational intelligence algorithms (fuzzy logic and genetic) applications and security issues (intrusion detection and prevention) over network.